RLG DigiNews
August 15, 1999, Volume 3, Number 4
ISSN 1093-5371

Table of Contents

Feature Articles

Illustrated Book Study: Digital Conversion Requirements of Printed Illustrations

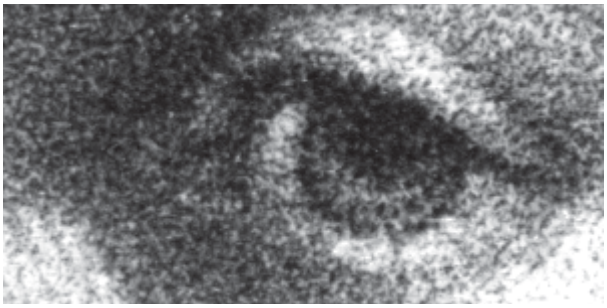
Anne R. Kenney, Cornell University Library
ark3@cornell.edu
and
Louis Sharpe II, President, Picture Elements, Inc.
lsharpe@picturel.com

Cornell University Library, Department of Preservation and Conservation, and Picture Elements, Inc. have issued their final report to the Library of Congress Preservation Directorate on their joint study to determine the best means for digitizing book illustrations. The project focused on illustration processes prevalent in 19th and early 20th century commercial publications, and was intended as a first step in the development of automated means for detecting, identifying, and treating book illustrations as part of a production digital conversion project.

Characterizing Illustration Attributes at Various Levels
A representative sample of illustration processes were assembled from Cornell's circulating collection. These included wood and metal engravings, halftone, etching, photogravure, mezzotint, lithograph, and collotype. An Advisory Committee of Cornell and Library of Congress curators, faculty, and other experts in printmaking and the graphic arts characterized the key attributes of each sample illustration reviewed, and their descriptions have been summarized in Table 1 of the report. Committee members assessed the significant informational content that must be conveyed by an electronic surrogate to support various research needs. In making those assessments, the Advisory Committee was asked to reflect on the intended uses of the sample documents in the context of their having been issued as part of larger published works rather than as individual pieces of art.

The following three levels of presentation were determined:

structure: representing the process or technique used to create the original. The level required for a positive identification of the illustration type varies with the process used to create it. For instance, it is easy to make a positive identification of a woodcut or a halftone with the unaided eye. The telltale reticulation of a collotype, however, may only be observable at magnification rates above 25x.



Mezzotint at 940 dpi

detail: representing the smallest significant part typically observable close up or under slight magnification, e.g., two times. This view is based on the psycho-visual experience of the reader rather than any feature associated with the source document.

Mezzotint at 400 dpi

essence: representing what the unaided eye can detect at a normal reading distance, again a psycho-visual determination.

Mezzotint at 200 dpi

Mapping Illustration Process Types to Digital Content Types

Digital requirements to reflect the structure view were predicted by measuring the finest element of the various print processes, which was easy to do for those characterized by well defined, distinct edge-based features, including the engravings, the etching, and the halftone. Despite differences in their identifying characteristics, project staff measured features ranging from .02 mm to .06 mm in size, with the majority of them measuring .04 mm. Evidence of the collotype structure was found in microscopically thin reticulation lines, measuring .01 mm or finer. For those items that were continuous tone-like, exhibiting soft grainy, dotted, or pebbly structures (e.g., the photogravure, mezzotint, and lithograph), feature details were hard to characterize and measure. Feature size estimates ranged from .04 mm to below .01 mm.

Based on these feature sizes, we quickly concluded that the resolution required to faithfully represent the structural characteristics would overwhelm any scanning project involving commercially produced publications. At a minimum, the resolution needed to preserve structural evidence in the digital surrogate, calculated at one pixel/feature, ranged from 635 dpi to over 2,500 dpi.

Predictions of digital requirements for the essence view were based on visual perception under normal lighting at a standard reading distance of 16 inches. A person with 20/20 vision can expect to discern features as fine as 1/215th of an inch (118 micrometers). These human visual capabilities suggest that a reasonable digital requirement for an on-

screen view representing the essence of a page would be 215 dpi. Predictions of digital requirements for the detail view were pegged at 2x magnification, which would require a digital resolution of 430 dpi.

Digitizing Sample Pages
Each sample page was scanned at a variety of resolutions with 8-bit grayscale data captured. Grayscale data is essential to reproduce the subtleties of perceived tonality inherent in many of the illustration types. It also permits accurate representation of fully bitonal features (having little tonality) when the feature size decreases towards the size of the image sampling function. Grayscale images allow various techniques used by skilled illustration artisans to have the intended tonal effects. For example, grayscale can preserve the modulation of the acid bite in an etching or the variation of the depth of a gouge in an engraving. Grayscale further permits the production of reduced-resolution images from a high-resolution original by means of accurate scaling algorithms.

Evaluating Sample Images
A Web site was prepared containing sample images for the various levels of view. The essence and detail views were created from the high resolution images by a process of bi-cubic scaling. Project staff prepared two views of these images. View #1 presented image segments at their native resolutions in a 100% view (1:1). View #2 images were resampled up to 600 dpi using bi-cubic interpolation to allow reviewers to assess images that were the same size on screen.

The Advisory Committee met several times, both in Ithaca, NY and Washington, DC, and assessed the digital surrogates at the three levels of view, comparing them to the original illustrations with and without magnification, and to printouts created from the essence and detail images.

The Advisory Committee distinguished two meanings for structural representation. The first interpreted "structure" as a view that allowed for identification of process type; the second required a view that faithfully replicated the sample under review. The resolution demands for the latter are much higher. The Advisory Committee also noted that it may be difficult to differentiate between similar process types even at very high resolution without additional testimonial evidence conveyed by the original artifact. These include date of publication, creator's name, whether the illustration appears on a separate plate or paper stock, and whether there was evidence of a plate mark. Finally, committee members felt that process identification for the softer edged images required both close examination and a pull back view to reflect on the nature of the overall composition. For instance, identification of the lithograph process relied on assessing the crayon-like appearance of the representation as well as examination of the pebbly grain structure revealed at higher resolutions or under magnification.

In conclusion, the Advisory Committee determined that digital images could provide good (but not always conclusive) evidence of structure, at the price of very high-resolution image files. They concurred with project staff that while this might be justified for individual artwork or selective samples, this was an impractical expectation in digitizing most commercially produced monographs and journals.

Advisory Committee members generally agreed that the 400 dpi on-screen view sufficiently captured the detail present in the original when viewed close up or under slight magnification. The committee's judgment regarding the on-screen detail view was remarkably consistent, and varied little with the illustration type. The Committee concluded that 400 dpi, 8 bit capture represented a good cost-benefit requirement for imaging when process identification was not an absolute requirement. The value of this approach is that it represents an assessment of close reading requirements that are based on visual perception, not on the informational content of the original materials. This is an important distinction, and suggests a uniform approach to determining conversion requirements for items that contain a broad range of illustration types, or that are difficult to quantify objectively. It also represents a reasonable conversion requirement for mixed items, containing both illustrations and text. The complete work can be imaged at the same level; and files post-processed to reflect the best presentation of the informational content-on screen to support various views, or to be printed out to meet readers' needs, or to create an equivalent to a preservation photocopy or microfilm. Where analyzing the print process of the original source is critical to an understanding of the work, the Committee concluded that the artifact itself should be preserved.

There was broad consensus from the Advisory Committee on the adequacy of the 200 dpi on-screen view to represent

the essence of the original. Lower resolution versions - on the order of 70-100 dpi - will provide a fair likeness of the general image content of the original, but will not match the psycho-visual perception of the original at normal viewing distances. Some tradeoff of perception, however, may be justified in cases where the original can be viewed completely on-screen, particularly for users with lower resolution monitors. For instance, a reader could display the complete image at 200 dpi on an 800 x 600 monitor, only if the dimensions of the original illustration did not exceed 4 inches by 3 inches. At 100 dpi, the complete image could be displayed for illustrations whose dimensions did not exceed 8 inches by 6 inches. In the future, as monitor resolutions increase, the 200 dpi view may become a practical standard for presenting the essence of original graphic illustrations.

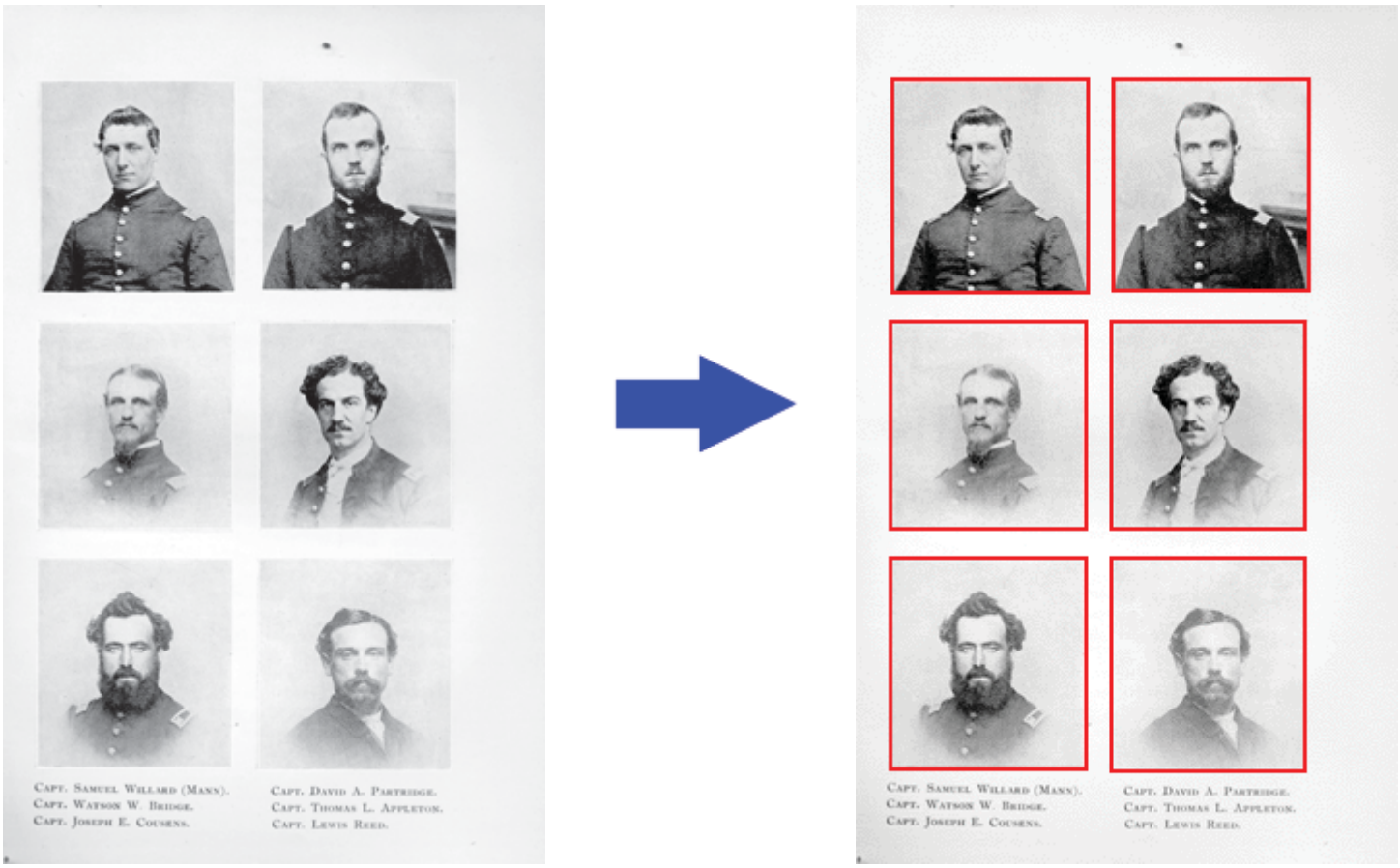
An Example Utility for Halftone Processing

A second phase of this project was devoted to post-capture processing of the raw image files for access. The focus of this effort centered on halftones, which are particularly difficult to represent in digital form, as the screen of the halftone and the grid comprising the digital image often conflict with one another. This can result in distorted image files exhibiting moiré patterns at various scales on computer screens or in printouts. A method for satisfactorily converting and processing halftones has been most pressing, as the halftone letterpress process became one of the most dominant illustration types used in commercial book runs beginning in the 1880s.

This project has resulted in the development of a practical, working utility to detect the location and characteristics of a halftone region on a page and appropriately process that halftone region independently from its surrounding text. Since this utility is not embedded inside a specific scanner, but runs externally on a UNIX server, it may be used on data from any scanner that can supply the appropriate raw bit stream (e.g., unprocessed grayscale of a sufficient spatial resolution). The source code and documentation is located at: http://www.picturel.com/halftone.

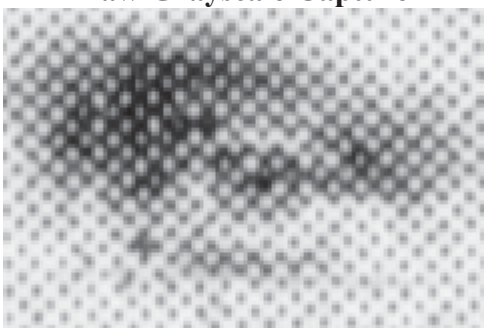
Below is an example of the utility locating the bounding rectangles of six different halftone regions on the same page, followed by an enlarged comparison of the unprocessed halftone.

Detecting Halftone Regions

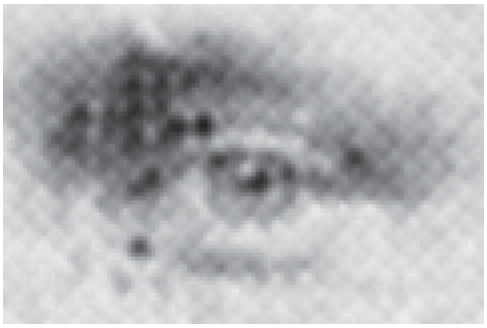


Processing Halftone Regions

Raw Grayscale Capture



Processed Halftone Information



Cornell project staff tested the prototype utility on a range of 19th and early 20th century halftones of various frequencies, from 110 line to 175 line screens. The following illustration demonstrates the halftone processing for a 120 lpi screen ruling.

120 lpi Halftone

Original Image

Processed Image

The utility worked equally well on halftones at other screen rulings placed at 45 degrees. Some Portable Document Format (PDF) files were prepared, which compare raw and processed halftone images next to one another, allowing easy experimentation with zoom levels and printing results.

Derivative Creation

Resampling halftone images introduces the likelihood of moiré patterning from screen frequency interference. This was evident when some of the full resolution images were scaled to 100 dpi to create derivatives for Web access. Obviously image processing routines can be used to minimize the introduction of moiré patterns as derivative images are prepared, but typical processes use blurring filters, which do not discriminate between screen rulings, and the results can vary dramatically. Rather than simply blurring, the utility's descreen process attempts to filter out the dominant frequency of the halftone screen. This is done by cascading a low-pass filter with a high-frequency emphasis filter, lessening the blurring effect. For most of the halftones, the descreening algorithm produced images that can be sub-sampled at any frequency without moiré patterns. Note the comparison of 100 dpi scaled images below. The one on the left was resampled without using a blur filter; the one in the middle was created using the standard blur filter; and the one on the right was created using the halftone utility.

DERIVATIVE IMAGES

Resampled from original halftone image

Resampled from a blurred halftone image

Resampled from an HPU processed image

Compound Documents

The practical problem inherent in digitizing and presenting halftones argues strongly for the application of the processing utility to scanned halftones. This leads, however, to a new technical problem - how best to re-aggregate this distinct grayscale image with the balance of the content from the enclosing page. In this project, we experimented with Adobe's Portable Document Format (PDF) because of its widespread use.

PDF permits multiple pieces of image content of varying types to be placed accurately onto an enclosing page. This allows the descreened halftone to be presented as a JPEG grayscale image while the textual portions of the page are thresholded to bitonal and compressed using ITU Group 4. Substantial space savings are achieved in this way.

Using another utility Picture Elements has created, the original page can be recomposed by laying the grayscale of the descreened halftone region on top of the bitonal text and white background and storing the result as a page in a PDF file.

As in any software project, the designers kept a wish list of ways in which the halftone processing utility could be improved. Since it is offered as public domain source code under the BookTools Project, others may undertake these enhancements and contribute the resulting improvements back to the community.

Conclusion

This study has produced a number of important results. The means for characterizing the key features of book illustrations as they pertain to digital imaging have been developed, and guidelines for assessing conversion requirements recommended. The basic groundwork for preparing an automated method for detecting and processing different illustration types has been prepared, and an example utility for processing halftones developed and tested. The halftone processing utility in particular will be a most welcome addition in the digital tool kit.

This project also facilitates a shift in thinking about how to create the highest possible image quality in a digital production project where cost and speed are of equal concern. This new capture architecture has the appropriate raw grayscale or color data collected from any scanner whose document handling capabilities suit the peculiarities of a particular item, such as a bound volume, a 35 mm slide, or a 40 inch wide architectural drawing. The scanner choice can be made solely on the basis of its physical suitability and the quality of its raw image data. All special processing and manipulation of raw data from these various sources is then performed in an off-line, largely scanner-independent manner by a centralized server we might call a post-processing server. In this way, we are not constrained by the variable and inconsistent processing offered within the many different scanners that are needed to overcome the physical peculiarities of each item in a collection. This work will be particularly important in developing the means for capturing bound volumes without having to resort to disbinding or to the creation of photo-intermediates.

Digitisation of Early Journals

Thaddeus Lipinski, Applications Manager, Bodleian Library, University of Oxford
tsl@bodley.ox.ac.uk

Introduction
The Internet Library of Early Journals (ILEJ) project was a three-year collaboration of the research libraries of the Universities of Birmingham, Leeds, Manchester and Oxford, funded by the eLib (Electronic Libraries) Programme. It ended in August 1998, with a final report published in March 1999. The project aimed to digitise substantial runs of pre-selected 18th- and 19th-century British journals, and to make these images available free of charge to the academic community. In addition, it explored variables in the digitisation, retrieval, and display processes, and evaluated the end users.

Another goal was to create a process for high volume, high throughput, low cost image production. This excluded labour intensive operations such as proofing OCR. Images were scanned from original volumes and from microfilm. OCR was used only for searching and indexing, not for displaying. In addition to OCR, a variety of sources were used to index the images, such as copy typed indexes and contents pages and existing electronic indexes supplied via third parties. To overcome the limitations of uncorrected OCR, the Excalibur EFS database was selected to provide a full-text fuzzy searching capability. Finally, the full-text database containing all the various elements and hyperlinks to the images were stored in SGML format.

Today, ILEJ supports a Web-based service of three 18th- and three 19th-century journals, together with indexes to the images. The journals are *Gentleman's Magazine*, *The Annual Register* and *Philosophical Transactions of the Royal Society* from the 18th century and *Notes and Queries*, *Blackwood's Edinburgh Magazine* and *The Builder* from the 19th century. The six titles display a diversity of typefaces, print and paper quality, article content and formats, and page size. Twenty year runs of the journals are available, yielding a resource of 110,000 images.

Scanning Issues
The journals selected for ILEJ are all out of copyright, so copyright considerations do not apply to their scanning and dissemination. In principle, the addition of 20th century journals to modern editions would need to take account of such issues, though the reverse has been true; commercial publishers with rights to the modern texts have discussed the possibilities of collaboration with the ILEJ consortium.

The ILEJ project concentrated on non-destructive, access quality scanning from both paper and microfilm originals. In 1996, the choice of scanning hardware was limited, and flatbed scanning from hard-copy originals (whether dismembered volumes or not) had been rejected at the outset. Several attributes were sought for paper scanning: overhead/cradle operation, protection of bound and fragile originals, speed, software correction for defects in originals (e.g., page curvature), high image quality, and greyscale scanning for output to a PC. These requirements limited the choice of scanner. At the time, the project compromised on the Minolta PS3000P open book scanner, although it then did not output greyscale images to a PC, nor can it handle resolutions beyond 400 dpi. This will limit the scope of ILEJ electronic masters in the future.

Similar difficulties with greyscale arose for scanning microfilm. The Mekel M500XL-G microfilm scanner used during the project can produce greyscale output, but again it is limited to a maximum of 400 dpi, with a 200 dpi maximum for greyscale TIFF images. Each frame of the microfilm showed two facing pages, and the scanning process converted each frame into two digitised images.

It has been difficult to get good results with the Minolta. Overall, inconsistencies in image quality can be considered inherent in the use of an open book scanner with this type of material. Throughput was on average 80-100 pages an hour, though production varied considerably from volume to volume. The rigid binding made pages undulate when the volumes were opened, causing shadowing in the gutter. Page curvature increased the shadowing. Many of the problems were caused by the variability of the original pages. Inconsistent text presentation (e.g., dense printing on one page, lighter printing on the facing page) proved troublesome for bitonal scanning and OCR processing. This was a particular problem with eighteenth century journals. Other problems that were exacerbated by bitonal scanning included show-through, foxing, and page discolouration. With open book scanning, the volumes can be placed anywhere within the Minolta scanning area, so the images had to be cropped to remove any extraneous black border.

With the Mekel microfilm scanner, a theoretical throughput rate of 600 frames per hour is possible, though the actual scanning rate was much lower. The effective throughput for *The Builder* and *Gentleman's Magazine* was 100 frames (200 images) per hour. Though this is twice as fast as that achieved with the open book scanner, the increased speed is offset by the need to manually reset the scanning parameters for almost every other page, and time taken to input metadata. Edge detection of the frame was problematic; the journal cover in the image created "false edges" that fooled the Mekel into displaying half a frame, or not winding the microfilm to the next frame. The microfilm images gave exactly the same problems as those presented by bound volumes for the same reasons (foxing, etc.) mentioned above. Though most images are legible, the quality is variable and in some cases poor. These problems led to many pages of *Gentlemen's Magazine* and *The Builder* being scanned in greyscale.

The problems with scanning and OCR are a direct result of the decision to preserve the original document. Dismembering the volumes and feeding the sheets through a suitable flatbed scanner would have solved many of the problems of scanning. Using a new generation open book scanner such as the Zeutschel may solve the problems associated with the Minolta.

Optical Character Recognition

Notes and Queries consists of short pieces on a wide variety of subjects, with lists of births, marriages and deaths, book reviews, and anecdotal information. This content is prohibitively expensive to index yet free-text searching would represent immense added value. An underlying assumption of ILEJ is that high (but not 100%) retrieval rates from a very large set of volumes is preferable to exact retrieval from a much smaller data set. Using uncorrected OCR and software to compensate for this would enable such a philosophy to be met.

Omnipage Pro Version 6.0 was used for text conversion as it provided good character recognition accuracy and allowed batch processing. However, the physical state of the material that caused problems in scanning also contributed to problems with OCR, *viz*. show-through and foxing. Repeating the OCR procedure with the same page sometimes gave different results.

Additional difficulties included small typeface of advertisements that cannot be OCRed effectively using 400 dpi images, and pages with complex structure including mixed fonts. In extreme cases, some pages could not be OCR processed, the software crashed for no discernible reason, or simply misread malformed text (e.g., *nice* becomes *miss*). Additionally, the OCR may result in incorrect translation because of:

- archaic fonts (recognition not possible, S translated as F), ligatures, and dipthongs
- archaic spellings (the dictionary recognises an archaic word as a modern word)
- passages in another language (Latin, Greek, French)

Other software should in principle improve a scanned image for text conversion. One such product is Sequioa Scanfix, which is used to despeckle and deskew images. Experiments showed these facilities had little overall effect on the accuracy of the OCR, and the software was used only to deskew images for display.

On typical machine produced documents, the leading OCR engines currently generate about 98-99.5% accuracy, a figure not obtained with all the ILEJ journals. Assessments of *Blackwood's Magazine* showed 98.5% accuracy, while *Notes and Queries* pages could be below 80%. Images from *Gentleman's Magazine* microfilming did not provide an acceptable level of OCR quality.

Fuzzy Searching

For two journals, ILEJ allows the options of simple searching of OCR or "fuzzy matching." Excalibur EFS document retrieval software is used to provide the latter facility. EFS uses fuzzy searching algorithms to compensate for uncorrected OCR. If an exact match cannot be found for a search term, a degree of "fuzziness" will attempt to find partial matches. For example, searching for "Manchester" may return "Manchester", "Manc~~ster", "Mansfield" or even "Worcester" if the degree of fuzziness is large. EFS ranks the hits so that the most likely (appropriate) matches appear first. Fuzzy matching radically increases the noise associated with hits. In addition, with the service version of EFS, boolean logic is restricted to 100% exact matching, so fuzzy matching cannot be used with multiple choices.

As an adjunct to the ILEJ system, the EFS server is inflexible to use, though the interface tries to hide this. It does not provide any bibliographic information with the text, other than its own internal index information. This may not be a problem as the displayed image (page) contains its own bibliography. Criticism of EFS must be balanced by the knowledge that it is properly intended to be used as a closed system for modern documents.

Image Display

A priority of the project was to provide a legible display. Ergonomically, scrolling of images larger than the display is preferably limited to one axis. It is more intuitive to scroll page images vertically rather than horizontally, so the images were sized with a fixed width bias where possible.

The original bitonal TIFF images were scaled to fit an 800 pixel wide screen, and grey was added to the resulting GIF images. This made the images more legible; as the type becomes smoother and any slightly broken character is filled in. The microfilm greyscale images were saved as JPEG files. Each page image was cropped to the text and a white margin was added later for clarity. Therefore, the displayed page is not a true facsimile of the printed page. Margins are representative only; the virtual page size may differ from the original. Some show-through may not be visible in the digitised image or may appear as blotches on the image. As images are scaled for viewing on screen, different journals have had different scaling applied, so comparisons about font clarity between journals cannot be made.

Metadata

Metadata for the ILEJ project are represented in SGML (Standard Generalised Markup Language) files. SGML-marked up text stores several types of *intellectual* metadata, including subject, author, and title indexes and OCR text and bibliographical information for each page. No *administrative* metadata (e.g., file resolutions, compression systems used, etc.) are included. Web search terms and Dublin Core records appear in the <META> HTML tag within the ILEJ home page to aid discovery by Web search engines.

An innovation of the ILEJ project was the transparent combination of distinct metadata from several sources. Bibliographical information was keyed by the scanner operator when scanning each page. Subject, author and title indexes were created by keyboarding the journal indexes or contents lists, or imported, with permission, from the electronic *Periodical Contents Index*.

SGML was preferred to a proprietary database as it allows various metadata to be integrated into a single structure. The hierarchical nature of SGML encoding mirrors the structure of the original journals. In addition, it is an ISO-approved format that can be migrated easily to new platforms in the future.

Every journal is represented by one SGML file. The SGML files describe the physical rather than the intellectual structure of each volume. Each journal page is represented in the file by a single <DIV> element. A unique SGML identifier links each page to the associated metadata, which includes a link to the image file and bibliographic information, and any subject, author or title indexes or OCR text. The bibliographic information provides descriptions of the pages, including those with no given page numbers, and distinguishes between the different categories of pages (title page, text, indexes). The results of text and index searches refer to *physical pages*, not articles that may span pages. Searching differs in this respect from modern electronic literature, which generally allows *article-based* searching. With ILEJ, the results of OCR or index searching will not pick up phrases that are split between pages.

An intellectual hierarchy delimits the boundaries between distinct units such as weekly issues for *The Builder*, articles for *Philosophical Transactions*, or monthly issues for *Gentleman's Magazine*.

Open Text PAT version 5 search engine is used to access the contents of SGML files at Oxford University. The results are reformatted to HTML on the fly for viewing on the Web.

The richness of metadata enables complementary search strategies to be used within ILEJ. Using subject, author, and title indexes means that the listed search term is an exact replica of the original. However, no further analysis of the page is possible. OCR allows further searching, but noise and uncertainty will be inherent with the result. Using the bibliographic content, hierarchical browsing is possible. Many early journals exhibit erratic pagination sequences, and browsing allows pages with the same page numbers, jumps in page numbers or no pagination to be listed in the same order as the original journal.

Journal Use
Users of ILEJ range from linguists, historians, and sociologists through to scientists and engineers. *Gentleman's Magazine* (35%) and *Notes and Queries* (25%) are the most popular journals. In addition to academics and students, genealogists and individuals on private research also access the Web site. The births and deaths sections of the above journals are heavily used. Hydrologists and geologists also search the "newsy" publications, searching for past references to river flooding or earthquakes. Feedback is specifically requested from the users, so comments are received regularly. There are many requests for the run of journals or the range of titles to be extended.

Despite difficulties encountered in this project, the experience has been invaluable, and helps inform a subsequent project to digitize Broadside Ballads from the Bodleian Library.

Highlighted Web Site

Photographic and Imaging Manufacturers Association (PIMA)

Representing the interests of manufacturers of imaging products, PIMA is involved in developing ANSI and ISO standards for traditional and digital photography, and collecting and disseminating industry statistical information. Its Web site provides a wealth of information for those interested in digital photography, especially in the area of emerging standards. Information on the recent activities of the PIMA-sponsored Technical Committee on Electronic Still Picture Imaging (PIMA/IT10) is also available. For example, the Committee has recently been involved in the development of a new resolution standard (ISO/FDIS 12233, Photography - Electronic Still Picture Cameras Resolution Measurements), which includes a resolution test chart and three measurement methods based on this target.

FAQs

Question:
I have recently read about the Octavo Corporation product that publishes and preserves rare materials using digital tools. Can you tell me how decisions were reached regarding the desired image quality, what your technical requirements are, and what are the plans for archiving the files?

Answer:
We contacted Patrick Ames, the CEO of Octavo to respond to the question. Octavo Corporation is a small for-profit company, staffed by technologists and publishing personnel, whose charter is to build a marketplace for "digital rare books," and to develop the technology and processes for doing so. Octavo is very concerned with editorial issues such as text, translation, authorization, and provenance, and bibliography, and many of the final publications contain both text and images.

On the technical side, Octavo's CCD array digital camera back is attached to a standard 4x5 field camera which supports a variety of select large format lenses. The camera is capable of capturing 6,000 x 8,000-pixel digital images in 32 bit RGB color. These resolutions create source files that are saved as TIFF, and range from 140 Mb to 750 Mb.

The source material is archived on CD-ROM as simple EPS (Encapsulated Post Script) files, compressed via the EPS default compression scheme. EPS is used for its production efficiencies and its cross-platform capabilities. There are numerous automation scripts that Octavo has written for the EPS files to support production or archiving. Octavo

makes three copies of each book, retaining one for production, giving one to the sponsoring library or museum for their own use, and archiving one in a different geographical location. On the Octavo network there are tapes, magnetic optical drives, arrays, and other devices. By June of 1999, one and a half years into production, Octavo has amassed a terabyte of data.

Calendar of Events

Digital Resources for the Humanities 99
September 12-15, 1999
To be held at King's College, London, this conference is a forum for all those affected by the digitization of common cultural heritage resources. The presentations will include academic papers, panel discussions, technical reports, and software demonstrations.

Second Annual E-Book Workshop
September 21-22, 1999
The National Information Standards Organization, (NISO) and the National Institute of Standards and Technology (NIST) will jointly sponsor the Electronic-Book Workshop to be held in Gaithersburg, MD, on the NIST campus. The workshop will focus on the technology surrounding the emerging electronic book products: the standards, the content, and the applications. It will showcase the major companies and developers supporting the E-Book.

Third European Conference on Research and Advanced Technology for Digital Libraries
September 22 - 24. 1999
Paris, France will be the location of this conference, which has as its main objective the bringing together of researchers from multiple disciplines. They will present their work on emerging technologies for digital libraries. The conference provides an opportunity to develop a research community in Europe focusing on digital library development.

International Symposium on Digital Libraries 1999 (ISDL'99)
September 28-29, 1999
To be held at the University of Library and Information Science, Tsukuba, Japan, this symposium will provide an international forum for papers and discussions by researchers, developers, and practitioners working on digital libraries.

American Society for Information Science: Annual Meeting
October 31 - November 4, 1999
ASIS will be held in Washington, DC, and the theme this year is: Knowledge: Creation, Organization and Use. The conference will look at knowledge creation, acquisition, navigation, correlation, retrieval, management, and dissemination.

Announcements

Joint National Science Foundation (NSF)/ Joint Information Systems Committee (JISC): International Digital Libraries Initiative
Six jointly funded international digital library projects have been announced, including one from the University of Michigan/CURL entitled, "Emulation Options for Digital Preservation: Technology Emulation as a Method For Long-term Access and Preservation of Digital Resources."

For further details of the NSF/JISC joint program contact: Stephen M. Griffin, sgriffin@nsf.gov, or Norman Wiseman, head.programmes@jisc.ac.uk.

National Science Foundation Announces Awards for Digital Libraries Initiative - Phase 2
Five United States government agencies working in partnership have awarded funding for projects that have three major research components; Research, Testbeds, and Applications; Undergraduate Emphasis Components; and

International Digital Libraries Collaborative Research. Of the projects funded, a number will address digital preservation issues, including Cornell University's Project PRISM, a design system to ensure the integrity of digitized texts; University of Pennsylvania's project that will explore how to trace and record provenance data; and Michigan State University's project, which will investigate methods of preserving digital audio files.

InterPARES: A Project to Investigate Preservation of Electronic Records
Funded in part by the National Historical Publications and Records Commission, this research project will focus on the long-term preservation of vital organizational records and critical research data created or maintained in electronic systems. The InterPARES Project (International Research on Permanent Authentic Records in Electronic Systems), will investigate and develop theories, methodologies, and prototype systems required for the permanent preservation of electronic records.

Digital Culture: Maximising the Nation's Investment:
A Synthesis of JISC/NPO Studies on the Preservation of Electronic Materials
The National Preservation Office has made available this report that summarizes seven recent digital preservation research studies supported by the Joint Information Systems Committee of the Higher Education Funding Councils, and the National Preservation Office. Free copies can be obtained from: Julia Foster, julia.foster@mail.bl.uk.

Preserving the Whole: A Two-Track Approach to Rescuing Social Science Data and Metadata
This report from the Digital Library Federation explores options for salvaging quantitative data stored in technically obsolete formats and its associated documentation stored on paper.

Model Editions Partnership Prototypes Now Online
The Model Editions Partnership (MEP) prototypes for scholarly editions of historical documents are now available on the Web. The project's "Markup Guidelines for Documentary Editions" are also online. The Partnership is a consortium of seven documentary editing projects which includes: The Documentary History of the First Federal Congress; The Documentary History of the Constitution and the Bill of Rights; The Papers of General Nathanael Greene; The Papers of Henry Laurens; The Legal Papers of Abraham Lincoln; The Papers of Elizabeth Cady Stanton and Susan B. Anthony; and The Margaret Sanger Papers.

RLG News

RLG-DLF Task Force Addresses Long-Term Retention of Digital Information

In its continuing program to address the preservation needs of the research community, RLG has partnered with the Digital Library Federation, launching a task force to identify current practice for long-term retention of digital research resources.

To be completed in March 2000, this effort will first gather institutional digital archiving policies as well as documentation of the institutions' current digital archiving practices. Three working groups focusing on electronic/institutional records, locally digitized materials, and electronic publications will analyze these statements to determine where best practice is emerging and where more collaborative effort is needed to create best practice consensus. The task force will develop a policy/practice framework to communicate what it has learned, and the information will be made available on the RLG Web site.

The joint RLG-DLF task force builds on a strong base of RLG-sponsored work to remove obstacles to long-term retention of digital material. In 1994, with the Commission on Preservation and Access, RLG co-sponsored the Task Force on the Archiving of Digital Information. The task force report, co-edited by Don Waters and John Garrett, recommended work needed to address problems inherent in the preservation of digital materials. Colleagues around the world have taken up several of the recommendations. RLG's PRESERV members identified those most appropriate for RLG action. (See *RLG Preservation Working Group on Digital Archiving: Final Report*.)

Following one of the PRESERV working group's recommendations, RLG surveyed its members to take the pulse of digital archiving activity and problems. This research, done by Margaret Hedstrom and Sheon Montgomery from the

University of Michigan's School of Information, was published last year. (See *Digital Preservation Needs and Requirements in RLG Member Institutions*). One glaring problem highlighted by the report was the lack of digital archiving policy statements, let alone documentation of existing practice, in institutions mandated to preserve acquired or created digital resources for the long-term. But as the RLG survey documented, creating digital preservation policies is a difficult task. The lack of good models for digital preservation, together with uncertainty about the most appropriate methods and approaches, appear to be major obstacles to developing effective policies and practices. The RLG-DLF Task Force addresses this need.

The task force charge, list of participants, and timeline can be found at the RLG Web site.

For further information about the RLG-DLF Task Force, contact Robin.Dale@notes.rlg.org.

RLG Provides Access to AMICO Library

The Art Museum Image Consortium's AMICO Library™ of digitized works of art is available to universities, schools, museums, and public libraries for institution-wide access over the World Wide Web, through the Research Libraries Group's enhanced Eureka® search system. This unique database for teaching and research - presently 50,000 works, and growing - contains diverse forms of art, such as paintings, sculptures, prints, drawings, photographs, and decorative arts. These works come from around the world: North America, including pre-Columbian (Meso-American) art; Europe, including ancient Greece and Rome; Asia, including ancient Asia minor; Africa, including ancient Egypt; South America; and Oceania. The library also covers all periods, from the ancient world to contemporary art.

In order to provide research access to the database, RLG received the data from AMICO and enhanced RLG's Eureka interface to accommodate images, to provide links to rights information, and to offer functionality to support the use of the AMICO Library.

AMICO members provided catalog records, images, and in some cases, additional multimedia documentation to AMICO, where they were consolidated, normalized, and enhanced.

Records describing the works at the item level were delivered to RLG in the AMICO Data Dictionary format, and are the basis for retrieval and sorting of results.

The images AMICO members provided were 1024 by 768 pixel 24-bit color TIFF files. RLG derived JPEG images in four sizes for delivery through our web interface:

- Thumbnail (128 pixels maximum height and width, shown in brief multi-item displays)
- Snapshot (250 pixels maximum height and width, for display beside full textual data)
- Inspection (480 pixels maximum height, 640 pixels maximum width, for study and presentation)
- Presentation (768 pixels maximum height, 1024 pixels maximum width, for study and presentation)

Depending on the type of AMICO license agreement an institution signed, users may also order the TIFF files, which are made available via FTP. Image metadata is provided for each TIFF and derivative image.

RLG's Eureka search system provides:

- Easy yet powerful searching
- Choice of several sizes of images
- The ability to save identified works into a personal notebook
- Storage of notebook contents for later access plus the ability to share the notebook with others

Students, instructors, librarians, and cultural and art historians can look for images through a wide variety of approaches - by title, holding museum, subject, format, creator, and much more - and create subsets of information they need for local use. They can view images at various sizes and compare them side-by-side. And they can conduct collaborative or independent research in an interactive, reliable online environment.

For more information: www.rlg.org/amico and www.amico.org

Hotlinks Included in This Issue

Feature Articles
Broadside Ballads from the Bodleian Library: http://www.bodley.ox.ac.uk/ballads/
BookTools Project: http://www.picturel.com/booktools
Halftone Conversion Utility Tool (HCUT) source code & documentation: http://www.picturel.com/halftone
HCUT conversion samples: http://lcweb.loc.gov/preserv/rt/illbk/HCUT.htm
HCUT sample Portable Document Format (PDF) files: http://www.picturel.com/halftone
HCUT Wish List: http://www.library.cornell.edu/preservation/illbk/ibs.htm#2106
Illustrated Book Study (IBS) final report: http://lcweb.loc.gov/preserv/rt/illbk/ibs.htm
Illustrated Book Study sample images: http://www.library.cornell.edu/preservation/illbk/AdComm.htm
The Internet Library of Early Journals (ILEJ) final report: http://www.bodley.ox.ac.uk/ilej
Table 1 of the IBS report: http://lcweb.loc.gov/preserv/rt/illbk/ibs.htm#table1

Highlighted Web Sites
Photographic and Imaging Manufacturers Association (PIMA): http://www.pima.net/

FAQs
Octavo Corporation: http://www.octavo.com

Calendar of Events
American Society for Information Science: Annual Meeting: http://www.asis.org/Conferences/AM99/
Digital Resources for the Humanities 99: http://www.kcl.ac.uk/cch/drh
International Symposium on Digital Libraries 1999 (ISDL'99): http://www.DL.ulis.ac.jp/ISDL99/
Second Annual E-Book Workshop: http://www.nist.gov/ebook99
Third European Conference on Research and Advanced Technology for Digital Libraries: http://www-rocq.inria.fr/EuroDL99/

Announcements
InterPARES: A Project to Investigate Preservation of Electronic Records: http://is.gseis.ucla.edu/us-interpares/
National Science Foundation Announces Awards for Digital Libraries Initiative - Phase 2: http://www.dli2.nsf.gov/projects.html
Model Editions Partnership Prototypes Now Online: http://adh.sc.edu/mepinfo/mep-info.html
Preserving the Whole: A Two-Track Approach to Rescuing Social Science Data and Metadata: http://www.clir.org/pubs/reports/pub83/contents.html

RLG News
AMICO: http://www.amico.org
AMICO Data Dictionary: http://www.amico.org/docs/dataspec.html
The AMICO Library: http://www.rlg.org/amico/
RLG-DLF Task Force Information: http://www.rlg.org/preserv/digrlgdlf99.html
RLG Preservation Working Group on Digital Archiving: Final Report: http://www.rlg.org/preserv/archpre.html
Digital Preservation Needs and Requirements in RLG Member Institutions: http://www.rlg.org/preserv/digpres.html
Task Force on the Archiving of Digital Information Report: http://www.rlg.org/ArchTF/

Publishing Information

RLG DigiNews (ISSN 1093-5371) is a newsletter conceived by the members of the Research Libraries Group's PRESERV community. Funded in part by the Council on Library and Information Resources (CLIR), it is available internationally via the RLG PRESERV Web site (http://www.rlg.org/preserv/). It will be published six times in 1999.

RLG DigiNews is produced for the Research Libraries Group, Inc. (RLG) by the staff of the Department of Preservation and Conservation, Cornell University Library. Co-Editors, Anne R. Kenney and Oya Y. Rieger; Production Editor, Barbara Berger; Associate Editor, Robin Dale (RLG); Technical Support, Allen Quirk.

All links in this issue were confirmed accurate as of August 12, 1999.

Please send your comments and questions to preservation@cornell.edu.

CONTENTS | SEARCH | HOME

Trademarks, Copyright, & Permissions